

Multiple Alignment

Anders Gorm Pedersen
Molecular Evolution Group
Center for Biological Sequence Analysis
gorm@cbs.dtu.dk

CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS

Global alignment score: 374

	60	70	80	90	100	110
alpha	QVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHL					
:::
beta	KVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLCVLAHFF					
	60	70	80	90	100	110

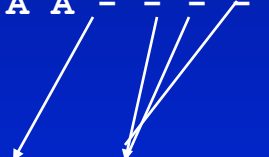
	120	130	140
alpha	PAEFTPAVHASLDKFLASVSTVLTSKYR		
	::: ::. .: .: .: .: .: .: .: .:		
beta	GKEFTPPVQAAYQKVVAGVANALAHKYH		
	120	130	140

Refresher: pairwise alignments

A	5							
R	-2	7						
N	-1	-1	7					
D	-2	-2	2	8				
C	-1	-4	-2	-4	13			
Q	-1	1	0	0	-3	7		
E	-1	0	0	2	-3	2	6	
G	0	-3	0	-1	-3	-2	-3	8
.								
.								
.								
	A	R	N	D	C	Q	E	G ...

- Alignment score is calculated from substitution matrix
- Identities on diagonal have high scores
- Similar amino acids have high scores
- Dissimilar amino acids have low (negative) scores

K L A A S V I L S D A L
K L A A - - - S D A L



$$-10 + 3 \times (-1) = -13$$

- Gaps penalized by gap-opening + gap elongation

Refresher: pairwise alignments

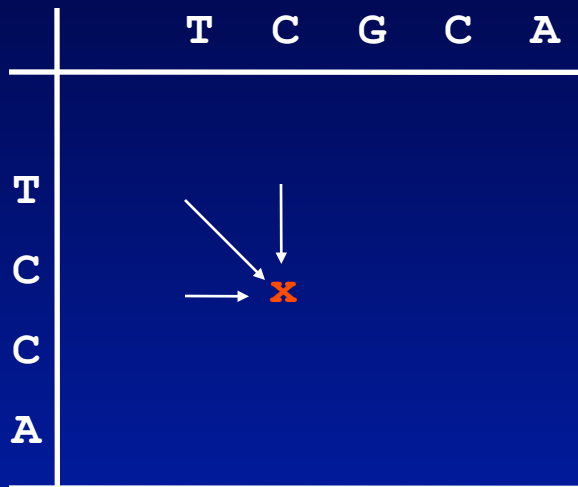
The number of possible pairwise alignments increases explosively with the length of the sequences:

Two protein sequences of length 100 amino acids can be aligned in approximately 10^{60} different ways



10^{60} bottles of beer would fill up our entire galaxy

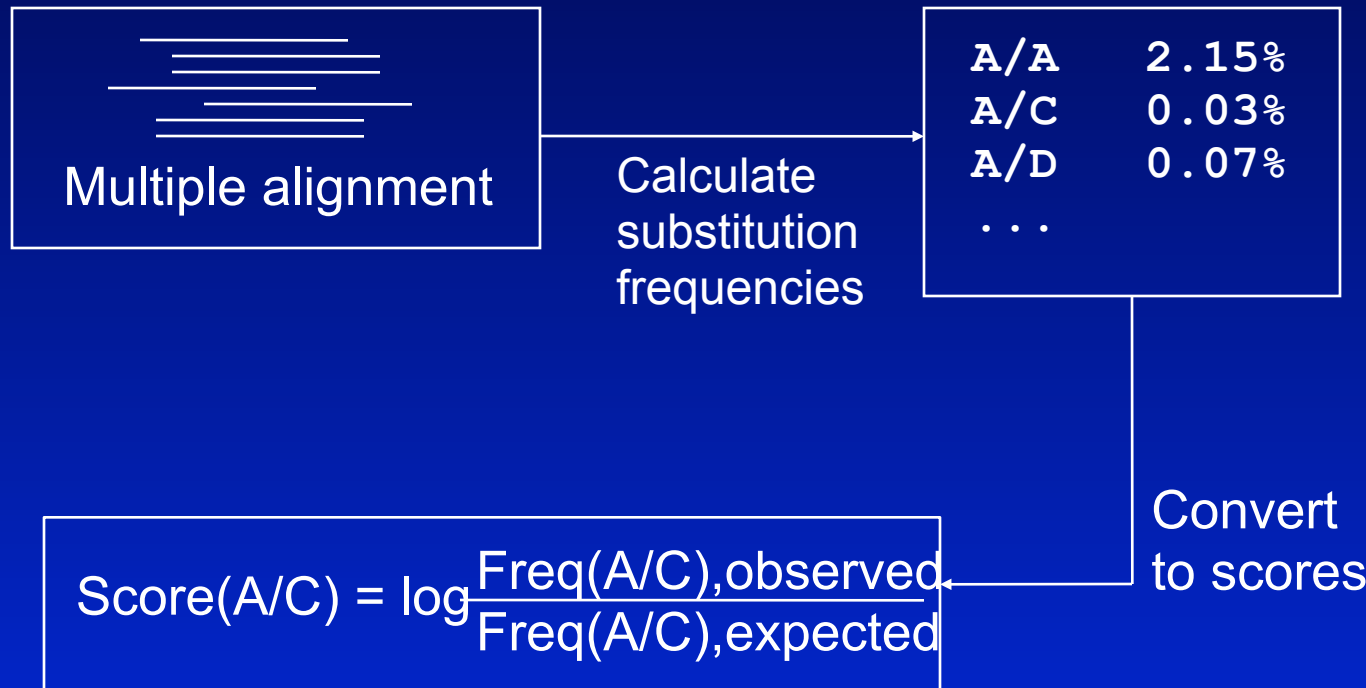
Refresher: pairwise alignments



- Solution:
dynamic programming
- Essentially:
the best path through any grid point in the alignment matrix must originate from one of three previous points
- Far fewer computations
- Best alignment guaranteed to be found

Refresher: pairwise alignments

- Most used substitution matrices are themselves derived empirically from simple multiple alignments



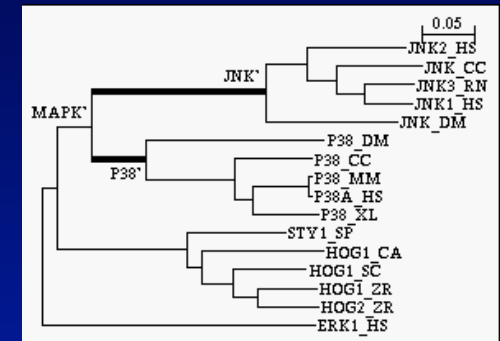
CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS



Multiple alignments: what use are they?

- Starting point for studies of molecular evolution

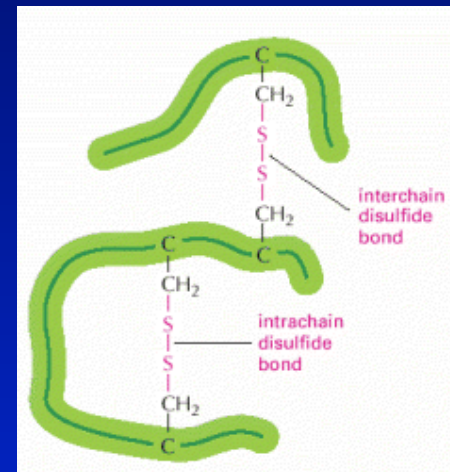
	***	*	*	****	*	**
af042103	AGATAGCTATAAAAATTAGGAGAACCAATTT	---	AAGAAAGAACAC			
af042105	GGATAGCTATAAACTTAGGAGAACCAATTT	---	AAGAAATAAAAC			
u16372	AGATAGCTGAAAAGTTAAGAGAACCAATTT	---	AATAAAAC			
u16374	AGATAGCTGAAAAGTTAAGAGAACCAATTT	---	AATAAAAC			
u16375	AGATAGCTGAAAAGTTAAGAGAACCAATTT	---	AATAAAAC			
u16373	AGATAGCTGAAAAGTTAAGAGAACCAATTT	---	GAGAAATAAAAC			
af042101	AGATAGCTGAAAAGTTAAGAGAACCAATTT	---	AAGAAATAAAAC			
u16376	AGATAGCTGAAAAGTTAAGAGAACCAATTT	---	AAGAAATAAAAC			
u16382	AGATAGCTGAAAAGTTAAGAGAACCAATTT	---	AAGAAATAAAAC			
u16381	AGATAGCTGAAAAGTTAAGAGAACCAATTT	---	AAGAAATAAAAC			
u16383	AGATAGCTGAAAAGTTAAGAGAACCAATTT	---	AAGAAATAAAAC			
u16385	AGATAGCTGAAAAGTTAAGAGAACCAATTT	---	AAGAAATAAAAC			
u16386	AGATAGCTGAAAAGTTAAGAGAACCAATTT	---	AAGAAATAAAAC			
u16377	AGATAGCTGAAAAGTTAAGAGAACCAATTT	---	AAGAAATAAAAC			
ruler	..360.....370.....380.....390.....4					



Multiple alignments: what use are they?

- Characterization of protein families:
 - Identification of conserved (functionally important) sequence regions
 - Construction of profiles for further database searching
 - Prediction of structural features (disulfide bonds, amphipathic alpha-helices, surface loops, etc.)

	100						105	
L	C	L	N	R	A	C	S	
M	C	S	N	Q	G	C	A	
A	C	G	S	S	A	C	N	
F	C	A	S	E	N	C	A	
T	C	D	S	N	G	C	Q	
M	C	R	L	R	D	C	S	



Scoring a multiple alignment: the “sum of pairs” score

...A...
...A...
...S...
...T...



One column
from alignment



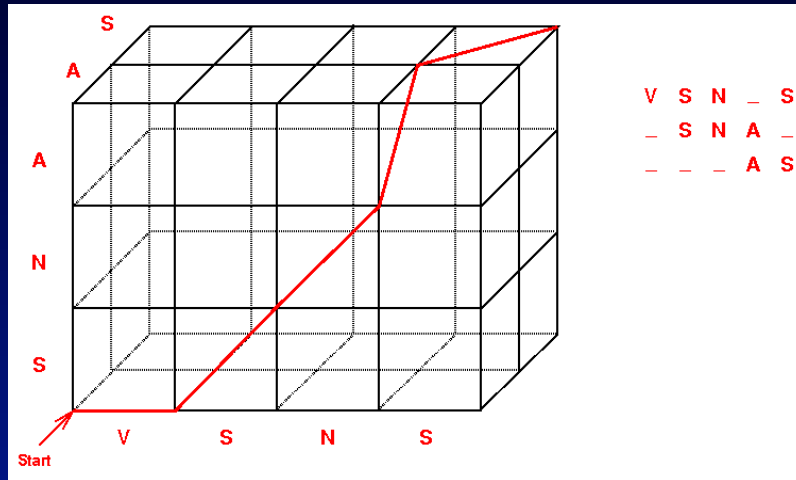
AA: 4, AS: 1, AT: 0
AS: 1, AT: 0
ST: 1

SP- score: $4+1+0+1+0+1 = 7$

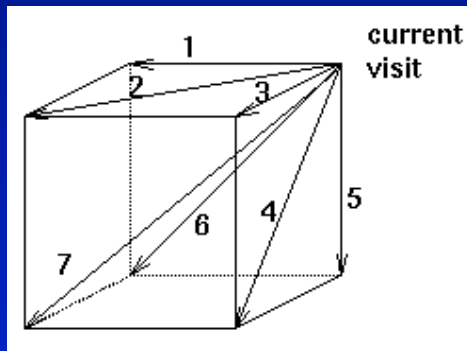
Weighted sum of pairs: each SP-score is multiplied by a weight reflecting the evolutionary distance (avoids undue influence on score by sets of very similar sequences)

=> In theory, it is possible to define an alignment score
for multiple alignments (there are several alternative scoring systems)

Multiple alignment: dynamic programming is only feasible for very small data sets



Dynamic programming matrix for 3 sequences



For 3 sequences, optimal path must come from one of 7 previous points

- In theory, optimal multiple alignment can be found by dynamic programming using a matrix with more dimensions (one dimension per sequence)
- BUT even with dynamic programming finding the optimal alignment very quickly becomes impossible due to the astronomical number of computations
- Full dynamic programming only possible for up to about 4-5 protein sequences of average length
- Even with heuristics, not feasible for more than 7-8 protein sequences
- Never used in practice

Multiple alignment: an approximate solution

- Progressive alignment (ClustalX and other programs):
 1. Perform all *pairwise* alignments; keep track of sequence similarities between all pairs of sequences (construct “distance matrix”)
 2. Align the most similar pair of sequences
 3. Progressively add sequences to the (constantly growing) multiple alignment in order of decreasing similarity.
- But important to realize almost no multiple alignment will be exactly right!